ASSOCIATION ANALYSIS OF UNRELATED INDIVIDUALS USING POLYMORPHIC GENETIC MARKERS

Jing Hua Zhao, Wendi Qian, University College London, UK and MRC Clinical Trials Unit, London, UK

BACKGROUND

Association analysis of unrelated individuals using multiple genetic markers are increasingly used. This could either be a marker-marker or marker-trait analysis. Haplotype phase uncertainty needs to be taken into account.

Clayton (2001) and Qin et al. (2002) have proposed heuristic EM and MCMC algorithms, but both are limited to SNPs. Here method in Clayton (2001) is extneded to multiallelic markers.

Previous global association tests using likelihoods do not give haplotype specific statistics, which are of considerable interest. We show via example they can be obtained during likelihood-based permutation tests.

METHOD AND IMPLEMENTATION

EXTENDING CLAYTON (2001)

- The new algorithm has the same feature of Clayton algorithm but considers multiple alleles when maintaining subject and haplotype lists.
- Appropriate procedure has been implemented to use results from multiple imputation as well as producing SAS programs containing the imputed data.

GLOBAL ASSOCIATION TESTS

- Likelihood-based permutation procedure is useful for producing LD-based statistics (Zhao et al. 1999)

1. MARKER-MARKER ANALYSIS

   x2 Statistic = -2(l[assuming association]-l[linkage equilibrium])

2. CASE-CONTROL ANALYSIS

   x2 Statistic = -2(l[case+control]-l[case]-l[control])

- Haplotype frequencies can be used for haplotype specific tests.

HAPLOTYPE SPECIFIC STATISTICS

- Simple Freeman-Turkey statistic for marker-marker analysis

$$FT = \sqrt{O} + \sqrt{O+1} - \sqrt{4E+1}$$

O, E = haplotype counts assuming linkage disequilibrium and linkage equilibrium

- Test of proportions for case-control heterogeneity analysis

$$z = \frac{\theta_1 - \theta_2}{\sqrt{V(\theta_1 - \theta_2)}}$$

θ_1, θ_2 =haplotype frequency parameter, *V*(.) = the variance function.

EXAMPLES

1. **HLA DRB, DQA and DQB markers (25,10,15 alleles) for 94 Schizophrenic patients and 177 controls. It shows the efficiency of polymorphic markers and use of haplotype specific tests.**

Table 1. Comparison of MCMC and EM estimates

Haplotype	Count	MCMC	EM	Eq	FT test	P value
22-2-12	62	11.4391	14.0193	0.4126	14.34	0.0020
4-8-1	62	11.4391	11.2545	0.5329	12.14	0.0020
9-4-1	46	8.4871	9.7786	0.3468	11.71	0.0020
1-1-7	41	7.5646	9.4095	0.2048	12.02	0.0020
6-5-3	34	6.2731	5.9737	0.2703	8.85	0.0020
8-5-3	27	4.9815	5.0698	0.1728	8.40	0.0020
14-8-2	20	3.6900	4.2366	0.1986	7.38	0.0020
6-5-2	18	3.3210	2.6979	0.3142	4.98	0.0020
17-3-13	13	2.3985	3.1291	0.0144	7.21	0.0020
10-7-6	12	2.2140	2.7675	0.0027	6.84	0.0020
21-1-9	10	1.8450	2.5830	0.0125	6.49	0.0020
18-2-14	9	1.6605	1.6605	0.0103	5.06	0.0020
3-1-7	8	1.4760	1.4760	0.0551	4.35	0.0020
9-4-4	8	1.4760	1.4760	0.067	4.26	0.0020
8-5-2	6	1.1070	1.3878	0.2009	3.35	0.0022
9-8-1	6	1.1070	<0.0001	0.5745	-2.67	N/A
12-5-4	6	1.1070	1.2915	0.0096	4.37	0.0020
16-8-2	6	1.1070	1.2915	0.0421	4.09	0.0020

Haplotype assignment by EM was unambiguous except for one individual with missing data.

Table 2. Comparison of individual haplotypes for HLA data

Haplotype	Case	Control	z-test	*P* value	Score test	*P* value
6-6-2	3.1915	0.0000	3.38	0.0003	2.95	0.0040
8-1-3	1.5957	8.2919	-3.13	0.0002	-3.11	0.0016
8-5-3	1.0638	7.2069	-3.10	0.0002	-3.05	0.0011
13-1-7	3.1915	0.2825	2.85	0.0001	2.89	0.0069
17-2-14	3.1915	0.5650	2.41	0.0008	2.45	0.0232
8-6-3	1.5957	0.0000	2.38	0.0012	2.40	0.0390
6-5-2	0.5319	3.8550	-2.27	0.0027	-2.14	0.0268
18-2-14	0.0000	2.5424	-2.20	0.0003	-2.21	0.0313
14-3-13	2.1277	0.2882	2.13	0.0023	1.38	0.3479
9-6-4	1.2395	0.0000	2.10	0.0014	1.96	0.1132
22-6-4	1.2413	0.0000	2.10	0.0008	1.96	0.1213
10-7-6	4.7872	1.6949	2.09	0.0025	2.14	0.0462
3-1-7	0.0000	2.2599	-2.08	0.0036	-2.08	0.0595
9-4-4	0.0000	2.2595	-2.08	0.0005	-2.08	0.0544

The z-statistic is comparable to score statistic, while empirical *P* values are due to different permutation procedures.

2. **ALDH2 markers and 130 alcoholic patients and 133 controls. This example shows the usefulness of LD-based analysis, the effect of missing data and importance of heuristic algorithm we implemented.**

Table 3. Eight ALDH2 region markers on Chromosome 12

Marker	Distance (b)	# alleles	# of missing individuals
D12S2070	> 450 000	8	251
D12S839	> 450 000	8	254
D12S821	~ 400 000	13	229
D12S1344	83 853	14	247
EXON12	0	2	261
EXON1	37 335	2	220
D12S2263	38 927	13	249
D12S1341	> 450 000	10	250

93 individuals with complete genotypes

- 1 month using only all markers by standard EM algorithm (Zhao et al. 2002)
- 6 days for 100 EM iterations using only possible haplotypes excluding two individuals with genotypes at only two loci
- 5 minutes for posterior trimming with threshold 0.001 but 8 hours with threshold 0.00001 (the new implementation)

3. **9 SNPs in APOC3/A4/A5 region from 3,012 individuals to study association with CHD and triglycerides. It shows drawbacks of heuristic algorithms and need to control for covariates.**

- Log-likelihoods by Qin et al. (2002), Clayton (2001), Zhao et al.(2002) were -13,988.0, -11,607.7 and -11,521.5, respectively, suggesting increasing optimality
- 30min for Qin et al. (2002) and Clayton (2001), but 5min by Zhao et al. (2002), so the raw sorting approach is less appealing, method using sufficient statistics is desirable.
- Method of Zhao et al. (2002) also gave equilibrium likelihood

CONCLUSION

- The heuristic EM and MCMC method is able to deal with multiple multiallelic markers, but it is still difficult to use it to obtain equilibrium likelihood and sufficient statistics are necessary for large sample.
- Haplotype specific statistics an be obtained from likelihood-based implementations. They are simpler than the score statistics.

ACKNOWLEDGEMENT

We are grateful of Dr padraig Wright for providing the HLA data, Prof Alun Thomas for his unpublished manuscript and result of using possible haplotypes on ALDH2 data. We thank Dr David Clayton and Prof Jurg Ott for making their programs available. Jing hua Zhao wishes to thank Colleagues in University College and Prof Pak Sham and other Colleagues in Institute of Psychiatry for collaborative work, NIA grant to Whitehall II study (AG13196) for support.

REFERENCES

Clayton (2001). **http://www-gene.cimr.cam.ac.uk/clayton/software**
Qin ZS. T Niu, JS Liu (2002). Am J Hum Genet 71, 1242-7
Zhao H, Pakstis AJ, Kidd JR, Kidd KK (1999). Ann Hum Genet 63:167-179, 1999.
Zhao JH, S Lissarrague, L Essioux, PC Sham (2002). Bioinformatics 18, 1694-5